





Customer



Industry
Information technology

Location Atlanta, Georgia

Business needs

To meet the growing demands of enterprise customers, Invoke required a cost-effective, high-performance platform that could support the ongoing development of applications as well as open-source AI inference and training tools that are integral to Invoke's product. The development platform had to provide reliability, robust support and security.

Solution created

Invoke equipped its developers with Dell Pro Max PCs* running NVIDIA AI Workbench so they'd have the on-demand, in-house resources needed for internal development and for training model components that power Invoke's core platform to ensure optimal performance and reliability.

Business results



Accelerates development efficiency by limiting cost overruns from cloud computing



Empowers AI developers to work anywhere with just a laptop



Improves research potential, customer service and security

Supporting technologies

- Dell Al Factory with NVIDIA
- <u>Dell Precision AI</u> 7780 mobile workstations with NVIDIA RTX™ Ada Generation GPUs and <u>NVIDIA AI</u> <u>Workbench</u>

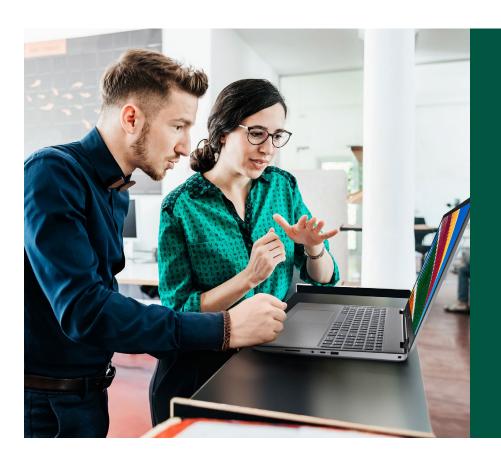
Several years ago, Kent Keirsey and colleagues created Invoke, an open-source Generative AI (GenAI) platform designed to support a wide range of industries including video games, film, TV, architecture, product design and marketing. Built on the Stable Diffusion text-to-image model, Invoke quickly gained traction due to its versatility, enabling professionals to generate creative content and designs. As the open-source platform grew, enterprise users began requesting advanced features, including enhanced reliability, responsive support and robust security to protect their intellectual property (IP). To meet this growing demand, Keirsey led the expansion of Invoke into an enterprise-grade platform, ensuring that businesses could leverage its capabilities in a secure and supported environment.

Drives doubledigit gains in developer efficiency.

Cloud costs limited development

Because Invoke is an early-stage startup, its developers had to be scrappy and make the most of limited funds. So, initially, machine learning (ML) engineers used pay-per-use cloud GPUs and compute to develop, train and tune the models supporting Invoke's core solution. However, cloud expenses quickly restricted research, experimentation and innovation. "One of the sneaky costs of a lot of cloud providers is data egress," explains Kent Keirsey, founder and CEO of Invoke. "When you're constantly iterating and uploading and downloading 6GB, 10GB and 50GB models from the cloud — and paying for all that data traversing the pipes — ML engineers quickly feel the pain of those costs, blocking what they can do."

Waiting for model uploads and downloads also slowed developers' productivity. Using a standard home 10Mbps connection, it would take developers about 1.25 hours to upload a typical 5GB Stable Diffusion XL text-to-image generation model. Additionally, delays during cloud-based product demos impacted the flow of presentations, which became a challenge during meetings and events. These cloud-based limitations highlighted the need for a more reliable, on-premises solution for developers.



"The Dell Precision
7780 laptop is the
most powerful mobile
workstation I've ever used,
and it's definitely more
stable than cloud-based
compute. Out of the box,
it's capable of running
Al workloads that we'd
normally need a tower for."

Kent KeirseyFounder and CEO, Invoke



Going full throttle on innovation

When Invoke learned about the Dell Pro Max PCs in the Dell Al Factory with NVIDIA portfolio, the company shifted to a largely on-premises development strategy. Dell Pro Max PCs deliver the performance that Invoke's developers require for GenAl workloads, and they enable anywhere, anytime work. Additionally, like all hardware offerings in the Dell Al Factory, the workstations arrive pre-tuned and ready to run Al tools including one of Invoke's tools of choice, NVIDIA Al Workbench. "The Dell Precision 7780 is the most powerful mobile workstation I've ever used, and it's definitely more stable than cloud-based compute," says Keirsey. "Out of the box, it's capable of running Al workloads that we'd normally need a tower for."

Advancing the world of visual media

Today, Invoke's developers use Dell Pro Max laptops to develop open-source training solutions and test the inference capabilities of the internal AI models that power Invoke's platform. "We're realizing double-digit gains in development productivity with our Dell Pro Max engineered for AI because we've empowered individuals to do as much as they can do," Keirsey says.

By running Invoke's open-source solutions on Dell and NVIDIA infrastructure, enterprise ML teams can iterate, test, and experiment locally with their models and workflows, keeping their innovative models and IP secure. And when they're ready, the teams can seamlessly deploy the models into Invoke's Enterprise solution, which runs in the cloud. "In this age of AI, velocity is everything," says Keirsey. "Dell Pro Max PCs eliminate friction by delivering the raw power needed to manage GenAI workloads and ML research while providing the security and reliability that our enterprise customers demand."



Recoups multiple hours each week per employee.

Saving hours of time and thousands of dollars in costs

Egress fees are no longer an innovation blocker, and developers and other employees who facilitate customer demos and training sessions have recouped the time they used to spend managing and waiting for "With NVIDIA AI Workbench, we can bridge together all of our Dell Pro Max PCs to create a powerful ecosystem of local compute."

Kent KeirseyFounder and CEO. Invoke



"In this age of AI, velocity is everything. Dell Pro Max PCs eliminate friction by delivering the raw power needed to manage GenAI workloads ... while providing the security and reliability that our enterprise customers demand."

Kent KeirseyFounder and CEO, Invoke

back-and-forth data exchanges with the cloud. Every employee that uses a Dell Pro Max laptop is saving hours of time each week and \$2,000–\$3,000 in monthly cloud costs.

Additionally, now that developers have the hardware resources they need at their fingertips, they can also experiment with innovative training techniques. If developers require more CPU or GPU power for model training, they can access these resources from other on-site Dell Precision workstations to avoid paying extra costs. "With NVIDIA AI Workbench, we can bridge together all of our Dell Pro Max PCs to create a powerful ecosystem of local compute," says Keirsey. "Dell Technologies and NVIDIA have done a really good job in making their technologies work together seamlessly."

\$2-3k savings in monthly cloud costs per employee

Improving customer experiences

Today, Keirsey and his team use their Dell Pro Max laptops to demonstrate how fast and easy it is to create and refine digital images with Invoke. The workstations have also improved the quality of customer trainings and demos, as the stability and performance of running everything locally eliminates the delays and inefficiencies often encountered when relying on cloud services. "With our Dell Pro Max laptops, we can take the power of a tower anywhere," says Keirsey. "Having the level of performance needed to run our powerful GenAl tools directly on a laptop is a significant leap in innovation, especially when we're hands-on with customers."

Learn More About Dell Pro Max PCs.

Contact a Dell Technologies Solutions Expert.

